分配意图与上行间接互惠：来自行为与 ERP 的证据[1]*

王婷[1] 赵梁佛[2] 杨金朋[2] 张丹丹[2,3] 雷震[2,4]

（1 四川大学，商学院，成都 610065；
2 西南财经大学，中国行为经济与行为金融研究中心，成都 611130；
3 四川师范大学，脑与心理科学研究院，成都 610066；
4 西南财经大学，经济学院，成都 611130）

摘 要 在真实世界与实验室环境中，上行间接互惠均广泛存在，其超越了重复的、封闭的互惠体系，成为推动人类大规模合作与社会秩序扩展的重要力量。虽然这一社会现象吸引了大量学者的关注，但是，现有文献存在两大不足：一方面，忽略了上行间接互惠中人们到底是以什么作为评判善意与非善意的基准这一核心问题；另一方面，大部分上行间接互惠文献未意识到“B 接受到 A 的善意或非善意后，B 以同样方式对待第三方 C”可能被“收入效应”这一竞争性假说解释。本文假设：当人们得到高于社会均值的分配时，真人分配条件下给第三方的金额会高于社会均值，真人分配条件下比电脑分配条件下给第三方的金额更高；当人们得到低于社会均值的分配时，真人分配条件下给第三方的金额会低于社会均值，但是真人分配条件下比电脑分配条件下给第三方的金额更低。研究采用两阶段独裁者博弈范式，结合 ERP 技术从大脑神经活动层面寻找证据。行为和脑电结果均较好支持了本研究的假设。真人分配比电脑分配诱发更大的 N1 波幅；低于社会均值的分配比高于社会均值的分配诱发更大的 FRN；在真人条件下，接受低于社会均值的分配比高于社会均值的分配结果诱发更大的 P3 波幅，在电脑条件下，接受低于或者高于社会均值的分配结果所诱发的 P3 波幅没有显著差异。本结果支持了基于社会分配均值的上行间接互惠假说，为该研究领域提供了新的理论基础和实验证据。

关键词 上行间接互惠，社会规范，分配意图，事件相关电位

分类号 B845; R395

---

[1] 收稿日期：2024-5-6
*四川大学专职博士后研究基金（skbsh2024-51）、国家自然科学基金重点项目（72033006）资助
通讯作者：雷震，E-mail：leizhen@swufe.edu.cn

1 引言

互惠是促进人类合作的重要机制之一。直接互惠即以善意直接回报施惠者，以非善意直接回报非善意者，可以用来解释个体间长期的、有着多次直接接触的合作行为（Nowak & Sigmund, 2005）。但是，很多情况下施惠者与受惠者无法建立长期的、直接的联系，人们不能将善意或非善意直接回馈给行为的发出者，而是传递给不相关的第三方，这一现象被称为间接互惠（Liu et al., 2022）。间接互惠突破了直接互惠的封闭体系，是促进人类社会大规模合作的重要力量，甚至是哈耶克提出的社会自发秩序（Hayek, 1980）的重要组成部分，因此识别间接互惠的内在机制具有重要的理论和现实意义。

间接互惠分为上行间接互惠（upstream indirect reciprocity）和下行间接互惠（downstream indirect reciprocity）两种形式。在本文中，我们将聚焦上行间接互惠，原因是上行间接互惠与下行间接互惠相比，不包含与声誉机制相关的策略性动机和高阶信念，便于更直接地考察间接互惠的内在机制。所谓上行间接互惠，通常是指 B 接受到 A 的善意（非善意）后，B 以同样方式对待第三方 C（Nowak & Sigmund, 2005）。根据新古典经济学的经济人假设，在完全匿名的一次性博弈中，B 接受了 A 的（非）善意后，由于无法从重复博弈的合作中获益而不会以（非）善意对待 C，导致间接互惠的链条难以形成（Sigmund, 2010）。然而，上行间接互惠得到了大量行为研究的验证，并与真实世界中的观察一致，这引起了心理学、行为科学领域学者的极大关注（Engelmann & Fischbauer, 2009; Horita et al., 2016; Rutte & Taborsky, 2007; 孙熠譞等, 2022）。

现有实验文献通常采用两阶段独裁者博弈（dictator game, DG）来刻画上行间接互惠（Gray et al., 2014; Hu et al., 2018; Liu et al., 2022）：在第一阶段 DG 中，被试 A、B 随机配对，A、B 分别作为分配者和接受者，B 不能拒绝 A 提出的分配方案；在第二阶段 DG 中，被试 B、C 随机配对，B、C 分别作为分配者和接受者，C 不能拒绝 B 提出的分配方案。根据 DG 实验文献元分析，被试平均分配 28%给他人，而分配大于 50%的份额给他人的被试极其少见（Engel, 2011）。但令我们感到惊讶的是，现有文献（例如 Gray et al., 2014; Hu et al., 2018）几乎都将 A 分配 50%给 B（等分分配）作为 B 判断 A 善意与否的分水岭，但对为何采用等分分配却缺乏充分论证。

我们认为，识别什么是善意与非善意不仅是 B 传递善意的前提，更是分析间接互惠认知机制的逻辑起点。通常，一个人给另一个匿名接受者分配一个接近于社会均值（28%）的分配值不会让人感到惊讶，而向上或向下偏离社会均值的分配值则可能被人们解读为善意或

非善意，也就是说，人们对他人的行为做出善意与否解读更可能是相对于社会均值这一参照点而言的。据此我们给出如下定义：善意（非善意）分配是指一个人给匿名接受者分配高（低）于社会均值的分配值。基于此，我们得以将上行间接互惠的定义中 B 接受到 A 的善意（非善意）后向 C 传递善意（非善意）转化成**假说 1：人们得到的分配值高（低）于社会均值时，倾向于给第三方一个高（低）于社会均值的分配值。**具体到本研究中，第一阶段 DG 共 10 元钱，我们将低于均值 28%的 1 元和 2 元作为非善意分配，将高于均值的 4 元和 5 元作为善意分配。这里没有考察 0 元，是因为在 0 元右侧邻域消费者偏好表现为非连续性（Niemand et al., 2019），而且去掉 0 元可以使善意和非善意分配出现的次数相等。

进一步地，如果假说 1 成立，该结果还可能被另一竞争性假说——收入效应假说——解释，后者指当人们的收入更高时倾向于给他人分配更多。如此，被试给第三方的分配变化便不是由高（低）于社会均值的分配值所代表的善意（非善意）触发的，而是由不包含意图的收入效应触发的。

所谓分配意图是指个体在进行分配决策时从一系列可选项中主动选择（不）利于接受者的选项的主观意愿（Falk & Fishbacher, 2006）。Falk 等（2008）探讨了意图的重要性，实验中 A、B 分别作为月光博弈（moonlighting game）的分配者和响应者，当分配值是由真人 A 做出时，B 的返还额随分配值增加而增加，但当分配值是由电脑随机产生时，B 的返还额不随分配值增加而增加，两种条件间返还额的差值就代表了 B 对 A 意图的反应。据此，本研究有必要考察在收入效应被控制之后，是否还存在由意图触发的间接互惠。

根据分配意图的含义（Falk & Fishbacher, 2006）并沿用以往实验研究的做法（Blount, 1995; Charnes & Rabin, 2002; Falk et al., 2008; Stanca, 2010; Zhang et al., 2016; Hu et al., 2018），本研究设置了真人分配和电脑分配两种条件，用以刻画意图的影响。相对于低于社会均值的电脑分配，B 接收到高于社会均值的电脑分配时给 C 分配更多，这可由收入效应来解释；但是，相对于低于社会均值的真人分配，B 接收到高于社会均值的真人分配时，B 可能不仅感受到高收入，而且感受到来自真人 A 的善意，则 B 给 C 的分配值就有可能比在电脑分配的情形时更高，高出的部分正是本文要识别的间接互惠。反之亦反。于是，我们得到**假说 2：当人们得到的分配值高（低）于社会均值时，人们在真人分配比电脑分配时分给第三方的值更高（低）。**

除了从行为层面验证以上 2 个假说，本研究还希望能从大脑层面打开上行间接互惠决策的“黑箱”。据我们所知，目前仅有 1 项 fMRI 研究（Hu et al., 2018）探索了分配结果与意图对上行间接互惠影响的脑机制，该研究发现被试的前脑岛、背侧前扣带皮层以及双侧前额

叶在接受到大于等分的慷慨或小于等分的贪婪分配时（与接受到等分的公平分配时相比）具有显著的神经激活，而且被试在接受到真人大于等分的慷慨分配时相对于电脑慷慨分配时（与等分的公平分配相比）右侧颞顶联合区和下顶叶有更显著的激活。但是正如上文所说，Hu 等（2018）的研究将等分 50%作为判断善意与否的阈值是不合适的。

本研究采用高时间分辨率的事件相关电位技术（even-related potential, ERP）从时间进程角度揭示上行间接互惠决策过程中的脑活动变化。根据前人相关研究（Liu et al., 2022; Miraghaie et al., 2022; Moore et al., 2021），本研究重点关注 3 个 ERP 成分：P1/N1、反馈相关负波（feedback-related negativity, FRN）和 P3。

位于外侧枕区的 P1 和 N1 具有相似的心理表征，均受到早期视觉注意和选择性注意的调控（Herrmann & Knight, 2001; Luck, 2014）。近期一项关于社会决策情境如何影响合作行为的 ERP 研究发现，与无社会互动情境相比，真人社会互动情境诱发了波幅更大的 N1（Moore et al., 2021）。结合以上发现，本研究认为第一阶段 DG 的分配者为真人时，分配结果会引发被试更多的选择性注意，从而产生较大的 P1 或 N1 波幅，于是得到**预测 1：与电脑分配相比，真人分配会诱发被试更大的 P1/N1 波幅**。

FRN 是位于额中央区的出现在结果反馈后 200~250 ms 的负波（Ma et al., 2015; Hoy et al., 2021），它反映与个人利益损失相关的加工过程（Gehring & Willoughby, 2002），还与社会期望和社会规范违背有关，越是违背预期或社会规范的分配结果，越能诱发更大的 FRN 负波（Wu et al., 2011; Mayer et al., 2019）。DG 研究发现，不公平的分配结果比公平的结果诱发更大的 FRN（Zhong et al., 2019; Li et al., 2020）。结合以上研究发现，我们认为低于社会平均分配的分配值对被试可能是一种违背预期的结果，于是我们得到**预测 2：与高于社会均值的分配相比，人们得到低于社会均值的分配时会表现出更大的 FRN 波幅**。

P3 成分是在反馈结果呈现后 300~600 ms 的顶区正波（Boudreau et al., 2009; Ma & Hu, 2015; Gong et al., 2022; Liu et al., 2022）。经典 Oddball 范式研究发现，P3 是由出现概率较小的刺激诱发的（Duncan-Johnson & Donchin, 1977; Johnson & Donchin, 1980），P3 波幅的增大表征了未预期事件的发生（de Bruijin et al., 2007）。因此我们得到**预测 3a：与高于社会均值的分配相比，人们得到低于社会均值的分配时会诱发更大的 P3 波幅**。同时，已有研究表明 P3 还表征大脑对当前结果主观价值的评价，对结果的主观价值评价程度越高，P3 波幅越大（Gu et al., 2011）。在本研究中，真人分配结果暗含分配者善意（非善意）的意图，而电脑分配的结果是随机的，没有包含分配者的意图。因此，我们认为被试对真人分配的结果具有更高的主观价值评价，需要消耗更多的注意资源，进而我们得到**预测 3b：真人分配比电脑**

分配诱发更大的 P3 波幅。不仅如此，社会决策相关研究发现 P3 同时受到分配结果和社会互动情境交互的调控（Qu et al., 2013）。相应的，由于低于社会均值的真人分配包含他人非善意的分配意图，具有更强的主观价值，因此我们得到**预测 3c：在真人条件下，接受低于社会均值的分配比高于社会均值的分配结果诱发更大的 P3 波幅**。而当被试接受到电脑的分配结果时，由于低于社会均值的电脑分配没有暗含他人的非善意意图，因此**电脑条件下，接受低于或者高于社会均值的分配结果所诱发的 P3 波幅没有显著差异。**

本研究采用两阶段 DG 实验框架探讨上行间接互惠，首次将第一阶段 DG 中，高（低）于社会均值而非等分的分配视为（非）善意分配，在此基础上通过操控第一阶段 DG 的分配者角色（电脑或真人）来刻画分配者意图，并借助 ERP 技术考察分配结果与分配意图对上行间接互惠的决策行为和脑神经活动的影响。

2 方法

2.1 被试

本研究招募在校本科生共计 42 人，年龄 18~22 岁，平均年龄 20.05 ± 0.14 岁（均值±标准误）。其中，男生 21 人，平均年龄 20.05 ± 0.21 岁，女生 21 人，平均年龄 20.05 ± 0.21 岁。所有被试均为右利手，视力或者矫正视力正常，无精神病史，没有参加过脑电实验，均自愿参与实验并签署了知情同意书。本研究通过了西南财经大学中国行为经济与行为金融研究中心伦理委员会的审核批准。根据相关研究（Liu et al., 2022）报告的效应量最小值（$\eta_p^2 = 0.09$），使用 G*Power 3.1.9（Faul et al., 2009）进行样本量估计（$f = 0.31$, $\alpha = 0.05$, 方差分析：重复测量，被试内因素），需要 32 名被试即可达到 99%的统计检验力（power），因此本研究被试量符合要求。

2.2 实验过程和实验材料

本实验包括两个部分：标准 DG 行为实验、两阶段 DG 间接互惠脑电实验。

第一部分为标准 DG 行为实验。被试作为分配者在自己与随机匹配的匿名接受者之间分配 10 元。该任务可以达到三个目的：首先，检验本实验样本的分配行为是否与以往 DG 实验文献中被试的分配行为一致，进而判断本实验样本是否有偏。同时获取研究样本的平均分配值，便于检验本研究将 3 作为区隔善意与非善意的合理性；其次，被试作为分配者进行分配后，可提高他们对间接互惠第一阶段 DG 分配值的可信度；最后，可据此考察不同利他水

平被试间接互惠行为的差异。

第二部分为间接互惠实验，即本研究的主实验。采用 2（分配结果：低于社会均值分配 vs. 高于社会均值分配）× 2（分配意图：真人分配 vs. 电脑分配）的被试内实验设计。我们沿用经典的两阶段 DG 框架来刻画上行间接互惠，所有被试将完成多轮两阶段 DG，每一轮都将重新随机匹配游戏对象。实验过程如图 1A 所示：第一阶段 DG 中，分配者 A 与接受者 B 共同拥有 10 元，被试作为接受者 B 收到匿名分配者 A 的分配后不能拒绝该分配方案；第二阶段 DG 中，分配者 B 与接受者 C 共同拥有 10 元，被试作为分配者 B 决定分配方案，匿名接受者 C 不能拒绝该分配方案。根据前文假说，首先，本文希望考察当被试在第一阶段 DG 中接受到高于社会均值的分配比接受到低于社会均值的分配时，被试在第二阶段 DG 作为分配者是否分配得更高。特别的，被试在接受到高于社会均值的分配时第二阶段 DG 的分配值是否高于社会均值，接受到低于社会均值的分配时分配值是否低于社会均值。上述实验结果如若成立，不能排除是由收入效应导致的，就本文研究的间接互惠而言，更应重点考察电脑分配与真人分配之间分配值是否存在显著差异，即，相对于电脑分配而言，B 接收到高（低）于社会均值的真人分配时，B 给 C 的分配值更高（低），高（低）出的部分正是本文需要重点识别的间接互惠。

主实验包括真人分配间接互惠和电脑分配间接互惠两个阶段，两个阶段的顺序在被试间平衡。每个阶段 156 试次，共计 312 试次。在每个实验阶段中，被试接受 3 种分配提议：高于社会均值的分配（4/6 和 5/5 各 30 试次）；低于社会均值的分配（1/9 和 2/8 各 30 试次）；填充条件（0/10、3/7、6/4、7/3、8/2、9/1 各 6 试次）。实验时长约 45 分钟。

实验开始前被试需进行 10 轮练习，待完全理解实验内容后开始正式实验。如图 1 所示，每轮实验任务开始时，屏幕将呈现“真人”或者“电脑”两个字，以提示当前轮次是由真人还是电脑做出分配。随后，在注视点后呈现第一轮分配者分配给被试的金额，分配数值持续时间 1500 ms。间隔 500~800 ms 黑屏后，屏幕呈现第二阶段 DG 中与被试配对的接受者 C 的照片。之后被试需要在 5 s 内用鼠标点击屏幕上的数字键将 0~10 的金额分配给接受者 C，点击“确定”键提交。实验开始前被试被告知，实验后将随机抽取五个试次，并依照这五个试次中被试所做的分配决策给被试发放实验任务奖金，以此激励被试在实验过程中认真作答。实验结束后，被试费为实验任务奖金加上参与实验的固定收入 50 元，平均约为 100 元。

为增加实验的可信度，本研究从前人研究的面孔材料库（Xie et al., 2021）中选取 312 张标准化证件照，男女各半，作为第二阶段 DG 的真人面孔素材呈现给被试（见图 1），面孔素材的性别在每个条件内进行了平衡。实验开始前，主试向每位被试强调：在真人分配任

务的第一阶段 DG 中，每个分配提议均是由不同提议者 A 提出的，在第二阶段 DG 中，被试需要对每位接受者 C 做出分配。主试也会给每位被试拍照，告知被试这是用作实验素材呈现给其他实验参与者的。实验材料由 E-prime 软件（Version 3.0）呈现，屏幕背景为黑色，第一阶段 DG 分配数值为白色数字，字体为 Times New Roman，字号 100，视角为 2.3×3.0°，真人面孔为彩色图片，呈现在屏幕中间。

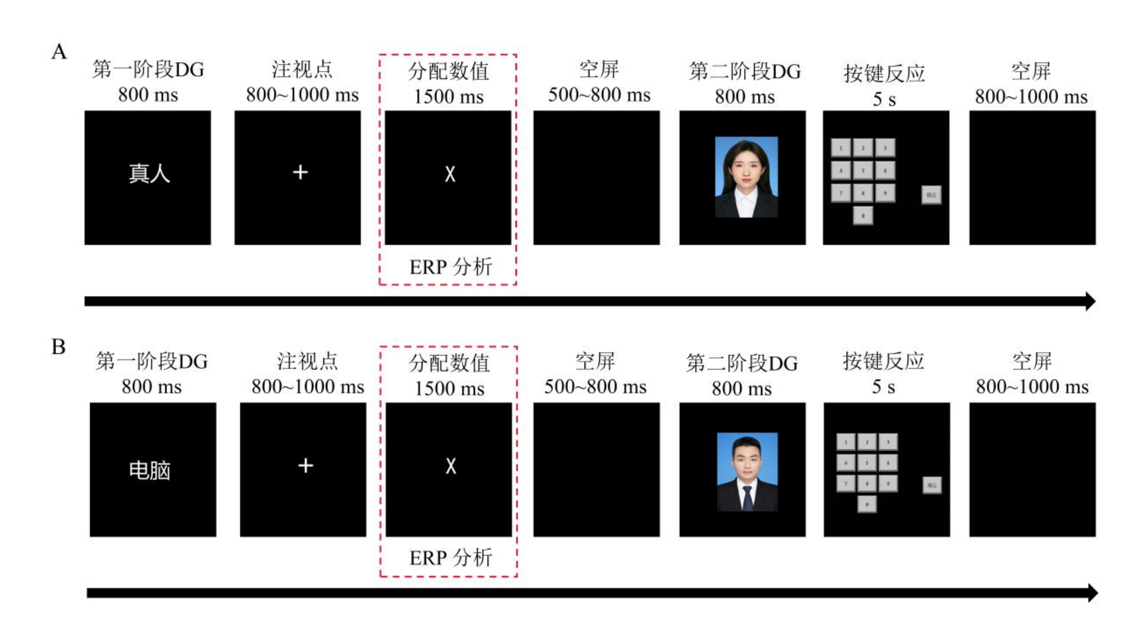


图 1 实验刺激材料。A, 单试次的真人分配间接互惠任务。B, 单试次的电脑分配间接互惠任务。为避免侵犯肖像权，此处使用课题组成员的照片作为示例。

2.3 脑电数据记录与分析

本实验采用德国 ANT Neuro 公司生产的 eegoTM mylab ERP 记录与分析系统，按照国际 10~20 系统扩展的 64 导脑电极帽记录 EEG。在线记录以 CPz 为参考电极，离线分析将参考电极转换为双侧乳突的平均电位。在被试左眼下侧粘贴电极记录眨眼引起的眼电。实验中所有电极与头皮间的阻抗值保持在 10 kΩ以下。A/D 采样频率为 1000 Hz。

离线数据分析使用基于 MATLAB R2017a（MathWorks）软件的 EEGlab 工具包，采样率降至 250 Hz，滤波参数为 0.1~30 Hz。使用独立成分分析的方法剔除眨眼伪迹，同时剔除幅度超过±100 μV 的试次。分段时间为第一阶段 DG 分配结果呈现前 200 ms 至结果呈现后 1500 ms，以提议结果呈现前 200 ms 作为基线对 ERP 进行矫正。

本研究感兴趣的 ERP 成分为枕区 P1/N1、前额 FRN 和顶区 P3。由于最终结果显示枕区 P1 在条件间无显著差异（差异开始出现的时间为 N1 时间窗），因此本研究统计了枕区 N1、

前额 FRN、顶区 P3 的平均波幅。用于测量三个 ERP 成分的电极点位置和时间窗已在数据分析前根据已有文献确定。N1 波幅测量采用外侧枕叶 PO5、PO6、PO7、PO8、O1、O2 电极点波形的均值，计算平均波幅的时间窗为 150~200 ms（Dong et al., 2010; Comesaña et al., 2013）。FRN 波幅测量采用前额中线附近 Fz、F1、F2、F3、F4、FCz、FC1、FC2、FC3、FC4 电极点波形的均值，计算平均波幅的时间窗为 200~250 ms（Zhang et al., 2013; Ma et al., 2015; Hoy et al., 2021）。P3 波幅测量采用顶区中线附近 Pz、P1、P2、P3、P4、CPz、CP1、CP2、POz 电极点波形的均值，计算平均波幅的时间窗为 300~600 ms（Ma et al., 2015; Wei & Zhou, 2020）。

行为数据和 ERP 成分幅度的统计分析使用 SPSS Statistics 23.0 软件, 分别进行 2（分配结果：低于社会均值分配 vs. 高于社会均值分配）× 2（分配意图：真人分配 vs. 电脑分配）重复测量方差分析。对显著的交互效应进行简单效应检验。描述性统计量均表示为“均值 ± 标准误”，显著性水平为 $p < 0.05$，采用η_p^2报告效应值。

3 结果

3.1 行为结果

3.1.1 标准独裁者博弈分配数值

被试在标准 DG 中平均将分配总额的 25.48 ± 2.73 %（2.548 ± 0.273 元）分给他人，如图 2A 所示，47.62%的被试分给他人 0~2 元，14.29%的被试分 3 元，38.1%的被试分 4 或者 5 元，没有人分配大于 5 元给他人。本实验的样本均值较好地代表社会平均分配值，t 检验表明本实验的分配均值(25.48%)与文献中的社会均值 28%统计上没有显著差异($p = 0.180$)，表明我们选取的样本具有代表性。同时，该结果支持了选取 1、2 元代表非善意分配以及 4、5 元代表善意分配的合理性。

3.1.2 假说检验 1

在真人分配前提下，**假说 1** 认为，当人们得到的分配值高（低）于社会均值时，倾向于给第三方一个高（低）于社会均值的分配值。*t* 检验结果发现，当人们接受到低于社会均值的分配数值时，更倾向于分配他人一个低于社会均值 25.48%的分配值（1.23 ± 0.18 元），$t = -7.33$，$p < 0.001$，$d = 1.131$；然而，当人们接受到高于社会均值的分配数值后，虽然分给他人的数值（2.37 ± 0.24 元）显著提高，$t = -8.57$，$p < 0.001$，$d = 1.322$（图 2B），但是总体上

并没有分给他人一个高于社会均值 25.48%的分配值，$t = -0.72$，$p = 0.473$。

那么，是哪些被试在接受到高于社会均值的分配时在第二阶段 DG 中分配给他人的金额没有显著高于社会均值呢？我们猜想那些低利他水平的人在面对他人的善意时，更不容易将善意传递出去。因此有必要对被试进行异质性分析。根据被试在标准 DG 中分配的数值对被试进行分类。我们将分配 0、1、2 元的被试纳入低利他水平组，而分配 3、4、5 元的被试纳入高利他水平组，从而分组考察不同利他水平被试的间接互惠行为的差异。低利他水平组被试有 20 人，含男生 11 人，平均年龄 19.85 ± 0.15 岁；高利他水平组被试有 22 人，含男生 10 人，平均年龄 20.23 ± 0.24 岁。

如图 2C 所示，与接受到低于社会均值的分配时相比（0.51 ± 0.12 元），低利他水平组被试虽然在接受到高于社会均值的分配时，对 C 的分配有了显著的提高（1.11 ± 0.18 元；$t = -4.65$，$p < 0.001$，$d = 1.04$），但是该分配结果仍显著小于社会均值 2.55 元（$t = -7.95$，$p < 0.001$，$d = 1.778$）。而高利他水平被试在接受到高于社会均值的分配时，分给 C 的数值不仅有了显著提高（3.52 ± 0.24 元；$t = -9.74$，$p < 0.001$，$d = 2.076$），更重要的该值显著高于社会均值（$t = 4.09$，$p = 0.001$，$d = 0.873$）。

以上结果表明，当收到低于社会均值的分配时，人们向第三方的分配也低于社会均值；当收到高于社会均值的分配时，具有高利他特征的人向第三方的分配高于社会均值，但具有低利他特征的人向第三方的分配不会高于社会均值。本结果**支持并进一步细化了假说 1**。

3.1.3 假说检验 2

假说 2 认为，当人们得到的分配值高于社会均值时，在真人分配比电脑分配时分给第三方的分配值更高；而当人们得到的分配值低于社会均值时，在真人分配比电脑分配时分给第三方的分配值更低。重复测量方差分析表明，分配结果的主效应显著，$F(1,41) = 53.15$，$p < 0.001$，$\eta_p^2 = 0.565$，被试接受到高于社会均值时，分给第三方的金额（2.29 ± 0.23 元）显著高于接受到低于社会均值分配条件（1.43 ± 0.19 元）。分配结果与分配意图的交互效应显著，$F(1,41) = 17.67$，$p < 0.001$，$\eta_p^2 = 0.301$。如图 2D，在接受低于社会均值的分配时，真人分配使被试分给他人的金额（1.23 ± 0.18 元）显著低于电脑分配时的分配金额（1.63 ± 0.22 元），$F(1,41) = 12.247$，$p = 0.001$，$\eta_p^2 = 0.230$。相反，在接受到高于社会均值的分配时，真人分配使被试分给他人的金额（2.37 ± 0.24 元）显著高于电脑分配时的金额（2.20 ± 0.22 元），$F(1,41) = 4.89$，$p = 0.033$，$\eta_p^2 = 0.106$。**假说 2 得到验证**。

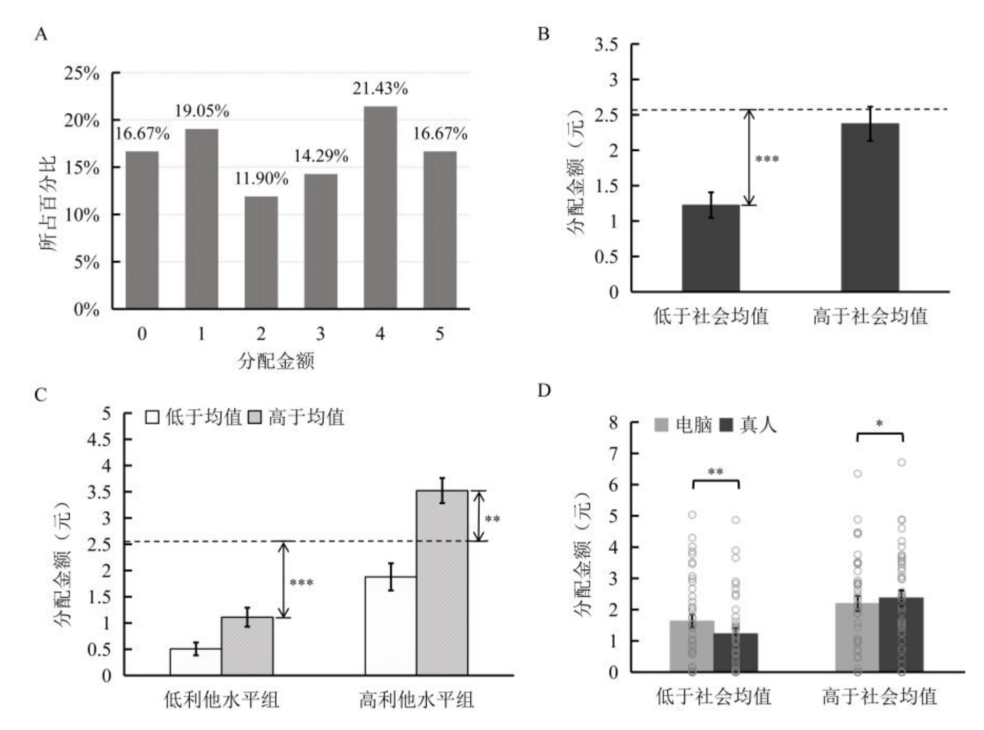


图 2　行为结果。A, 标准独裁者博弈分配数值。B, 被试在接受到真人的低于社会均值分配和高于社会均值分配时，分配给第三方的金额。C, 不同利他水平被试在接受到高（低）于社会均值的真人分配时，分配给第三方的金额。图中横虚线代表本研究的社会均值 2.548 元。D, 被试在四个实验条件下的分配金额。图中的误差条代表标准误。小圆圈表示单个被试的数据。*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$。

3.2 ERP 结果

3.2.1　N1 成分

如图 3 所示，分配结果的主效应不显著，$F(1,41) < 1$，高于社会均值的分配（-1.88 ± 0.38 μV）与低于社会均值的分配（-1.96 ± 0.40 μV）诱发的 N1 波幅没有显著差异。分配意图的主效应显著，$F(1,41) = 5.24$，$p = 0.027$，$\eta_p^2 = 0.113$，真人分配条件诱发的 N1 波幅（-2.26 ± 0.39 μV）显著大于电脑分配条件（-1.59 ± 0.42 μV）。分配结果与分配意图的交互效应不显著，$F(1,41) < 1$。**预测 1 得到验证**。

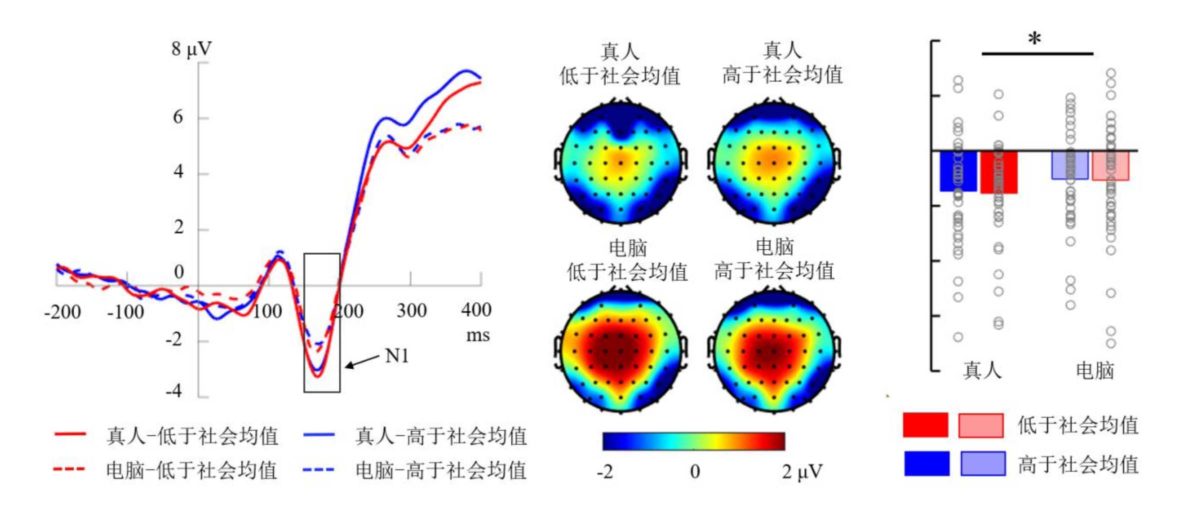


图 3　N1 的波形图、地形图和幅度值。波形图为 PO5、PO6、PO7、PO8、O1 和 O2 电极点的平均波幅。地形图为时间窗 150~200 ms 内的平均幅值。图中的误差条（error bar）代表标准误。小圆圈表示单个被试的数据。$^*p < 0.05$。

3.2.2　FRN 成分

如图 4 所示，分配结果的主效应显著，$F(1,41) = 38.68$，$p < 0.001$，$\eta_p^2 = 0.485$，低于社会均值的分配（1.67 ± 0.76 μV）比高于社会均值的分配（4.24 ± 0.70 μV）诱发更负的 FRN 波幅。真人分配条件（2.726 ± 0.826 μV）诱发的 FRN 波幅与电脑分配条件（3.19 ± 0.63 μV）诱发的 FRN 波幅没有显著差异，$F(1,41) = 1.19$，$p = 0.282$，$\eta_p^2 = 0.028$。分配结果与分配意图的交互效应不显著，$F(1,41) =1.26$，$p = 0.268$，$\eta_p^2 = 0.030$。**预测 2 得到支持**。

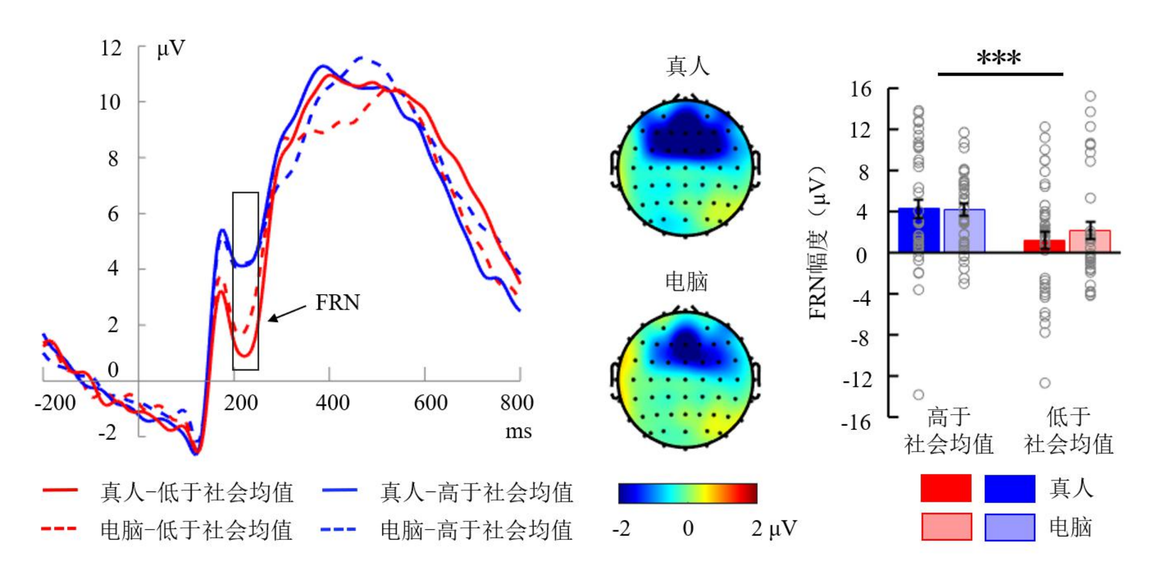


图 4　FRN 的波形图、地形图和幅度值。波形图为 Fz、F1、F2、F3、F4、FCz、FC1、FC2、FC3 和 FC4 电极点的平均波幅。地形图为时间窗 200~250 ms 内的平均幅值。为更直观反映 FRN 的地形图分布，此处为差值地形图：“低于社会均值条件”减去“高于社会均值条件”。$^{***}p < 0.001$。

3.2.3　P3 成分

如图 5 所示，分配结果的主效应显著，$F(1,41) = 25.14$，$p < 0.001$，$\eta_p^2 = 0.380$，即低于社会均值的分配（9.42 ± 0.74 μV）比高于社会均值的分配（7.93 ± 0.81 μV）诱发更大的 P3 波幅。分配意图的主效应显著，$F(1,41) = 51.33$，$p < 0.001$，$\eta_p^2 = 0.556$，真人分配条件诱发的 P3 波幅（10.83 ± 0.83 μV）显著大于电脑分配条件（6.53 ± 0.81 μV）。分配结果与分配意图的交互效应显著，$F(1,41) = 4.54$，$p = 0.039$，$\eta_p^2 = 0.100$，简单效应检验发现，在真人条件下，接受低于社会均值的分配结果（11.95 ± 0.83 μV）比接受高于社会均值的分配结果（9.70 ± 0.88 μV）诱发更大的 P3，$F(1,41) = 27.93$，$p < 0.001$，$\eta_p^2 = 0.405$。与此不同，在电脑条件下，接受低于（6.89 ± 0.77 μV）或高于社会均值的分配结果（6.17 ± 0.91 μV）在 P3 幅值上没有显著差异，$F(1,41) = 2.04$，$p = 0.161$，$\eta_p^2 = 0.047$。**预测 3 得到验证。**

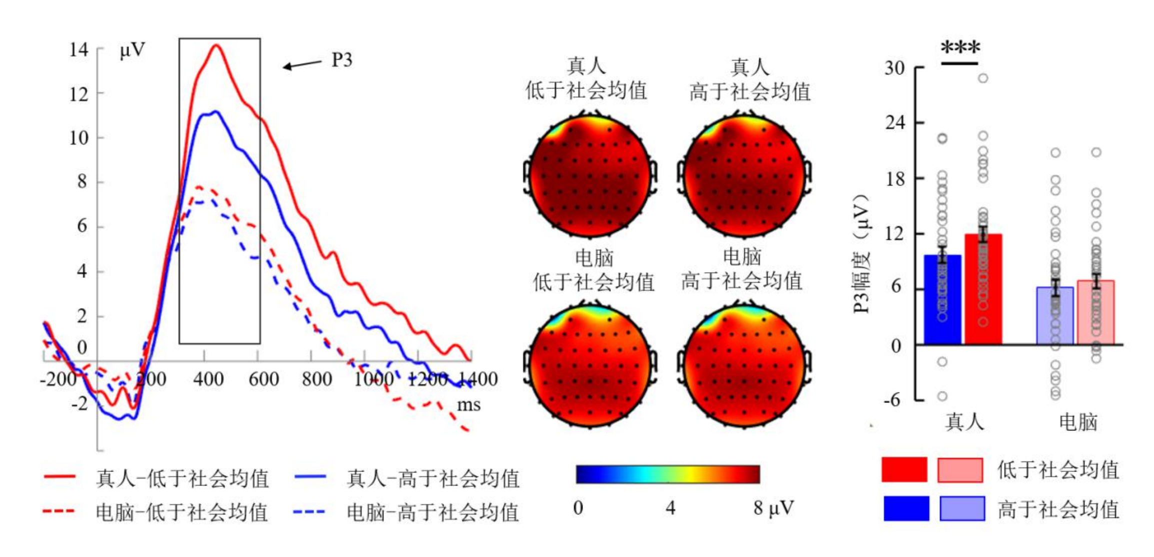


图 5　P3 的波形图、地形图和幅度值。波形图为 Pz、P1、P2、P3、P4、CPz、CP1、CP2 和 POz 电极点的平均波幅。地形图为时间窗 300~600 ms 内的平均幅值。***$p < 0.001$。

4 讨论

在真实世界中，“B 接受到 A 的（非）善意后，以（非）善意对待第三方 C”的上行间接互惠行为无处不在（Nowak & Sigmund, 2005），即使在排除了重复博弈的匿名和一次性博弈严苛的实验环境后，上行间接互惠仍然稳健地存在，这超越了新古典经济学的认识。更重要的是，上行间接互惠将封闭的互惠关系扩展到了第三方，这一特征具有重要的意义，如果说市场是哈耶克所提出的自发社会秩序扩展的外显力量，那么，上行间接互惠可以被视为这种秩序扩展的一种内在力量，促进了社会大规模的合作和人类社会的演化。学术界一直尝试

从不同角度探索上行间接互惠这一重要现象（孙熠譞等, 2022），但从本文的视角来看，现有文献对其内在机制的探索仍然不够。本文的主要创新在于选择了与以往将等分分配视为判断A是否善意的参照点（Gray et al., 2014; Hu et al., 2018）不同的研究框架，将A高于真实的社会均值的分配看成上行间接互惠中善意的分配，采用真人分配和电脑分配的两阶段DG进行行为层面的假说检验，并结合ERP技术从大脑神经活动进程层面寻找证据。实验结果从行为和脑神经两个层面支持了本文假说。

为什么我们要提出与等分分配作为A是否善意的参照点所不同的研究框架呢？首先，决策者的社会行为往往受到社会规范的影响（Kimbrough & Vostroknutov, 2016; Pereda et al., 2017; 黄馨茹等, 2023），间接互惠作为一种社会性行为，理应同样受到社会规范的影响。基于此，我们认为给定其他条件不变的情况下，识别A的分配是否善意不能离开社会规范而孤立存在。社会规范包括命令性规范（多数人赞成或不赞成的行为）和描述性规范（多数人实际采取的行为模式）(Cialdini et al., 1990)。命令性规范建立在“该怎样”的主观态度基础上，描述性规范建立在“是怎样”的客观分析基础上。由于前者的主观态度难以观测，在实验研究中较难获得准确的数据，而后者仅需收集每个个体的行为数据，便于进行实证检验，所以本文选择基于描述性规范进行分析。我们认为，个体与他人交往时不断更新对某社会行为的理解，大脑中逐渐形成基于自身经验的对某社会行为分布的信念，当遇到一个人的分配行为与（习得的）社会均值信念无异时不会感到惊讶，而向上或者向下大幅偏离该均值信念则有可能激发个体强烈的反应。个体遵从规范的意愿依赖于他/她关于他人遵从规范的信念（McBride & Ridinger, 2021），这是一种促进社会规范收敛的机制，人们关于社会均值的信念收敛到真实的社会均值可能是数学上的一个不动点，所以我们假设，就社会总体而言，人们关于社会均值的信念与真实的社会均值是无偏的。本文的实验结果支持了这一假说，我们发现当人们得到高于社会均值的分配时，分给他人的分配值更高（即将善意传递给他人），而当遭遇低于社会均值的分配后，分给他人的分配值更低（即将非善意传递给他人）。

现有文献主要将等分分配作为善意与非善意的分水岭，一个可能的原因是他们（不自觉地）选择了命令性规范作为判断的标准。例如在Gray等（2014）的行为实验中，实验者告知被试B，分配者A会将6美元中的0、3、6美元分给B，而B需要决定在第二阶段DG中作为分配者将6美元中的多少分给C。Gray等将等分分配作为公平，而将0和6分别作为贪婪和慷慨分配，此等分分配的依据来自另一项想象的分配实验（Messick & Schell, 1992）。后者仅考察被试作为旁观者的分配，这与自己参与的分配相去甚远，且不是一次性实验，难以排除重复博弈带来的对声誉的影响，我们认为Gray等（2014）以此实验的结论作为其研

究立论的依据有失严谨。又如，在 Hu 等（2018）的行为与 fMRI 实验中，在第一阶段 DG 设置等分、高于等分和低于等分试次各占 1/3，第二阶段 DG 中让分配者从不平等分配方案与等分方案中作选择，其中不平等方案根据 Fehr 和 Schmidt（1999）不平等厌恶（inequality aversion）假说生成。然而，根据 DG 实验文献元分析，几乎所有实验研究发现分配大于 50%的份额给他人的被试极其少见（Engel, 2011），同时在我们研究的标准 DG 实验中也没有人分配大于 5 元给接受者，见图 2A。Hu 等（2018）的实验中，被试接受到高于 50%的分配额的次数等于低于 50%的分配次数，这与被试在真实世界中难以遇到高于 50%的分配额的经验相冲突，实验结果将难以刻画人们面对真实世界中分配结果所做出的反应，所得到的间接互惠的结论也就难以通过外部有效性检验。即使将等分看成命令性规范，也需就等分行为在道德上的合理性进行调查，遗憾的是，Gray 等（2014）和 Hu 等（2018）并未收集这类数据，这样的实验设计难以排除实验者的需求效应（demand effect）。

本文另一创新点在于通过操纵分配者角色为电脑还是真人，考察了控制收入效应后因意图所触发的间接互惠是否存在。结果表明，与低于社会均值的电脑分配相比，低于社会均值的真人分配使被试 B 分配给第三方的数值更低，表现出更高的负向间接互惠。可能的解释是真人 A 低于社会均值的分配让被试感知到非善意的分配动机，从而产生 Nowak & Sigmund（1998）提到的消极情绪体验（例如愤怒），由于被试无法对非善意分配的施加者进行报复，于是出于宣泄消极情绪等目的，他们将非善意的分配传递给第三人。与之相反的，相较于真人 A 低于社会均值的分配，电脑低于社会均值的分配不涉及分配者的行为意图，于是被试 B 也就没有强烈的动机将非善意分配传递给他人。另一方面，本研究发现真人 A 高于社会均值的分配比电脑高于社会均值的分配使被试产生了更强的正向间接互惠。已有研究发现感激情绪或许是影响人们进行间接互惠传递的心理机制（Chang et al., 2012）。据此我们推测，当人们感知到高于社会均值的分配代表有意的善行而非随机事件时，人们对他人的分配可能产生感激的情绪（McCullough et al., 2001），从而更愿意将善意的分配传递给他人。Hu 等（2018）虽同样探讨了分配结果与分配意图对上行间接互惠行为的影响，但如前文所述，现有 DG 实验发现人们几乎不会将大于等分的分配给接受者（Engel, 2011），而 Hu 等（2018）将大于等分的分配方案的出现次数与小于等分的分配方案的出现次数设置为相等，或许会让被试在收到大于等分的真人分配方案时难以相信其真实性。更重要的是，当被试收到小于等分的贪婪分配时，由于贪婪分配包含本文所操纵的低于社会均值的分配和高于社会均值的分配这两种分配方案，使得被试既因收到低于社会均值的分配而倾向于给第三方一个较低的分配，又因收到高于社会均值而给他人一个较高的分配，两种相反的分配模式被归入同一分配条件来衡

量间接互惠，最终导致 Hu 等（2018）发现真人和电脑分配条件下人们的间接互惠分配在行为上没有显著差异，即在行为上基于意图的间接互惠假说未得到支持。

在脑电层面，本研究发现由 N1 表征的早期选择性注意受到分配意图的调控。真人分配比电脑分配诱发更大的 N1，这说明与电脑相比，真人分配获得被试更多的选择性注意。该结果**与预测 1 一致**。在漫长的人类进化史中，作为社会性动物的人类在交流互动中获得了竞争优势，并逐渐进化出对同类更为关注的特点。本研究的真人分配代表着人与人之间的真实社会互动情境，大脑在此情境下需要更强的注意资源投入（Lin et al., 2014; Moore et al., 2021），进而诱发更大的 N1 波幅。

实验结果还表明，FRN 波幅受到实际分配数值是否低于社会平均分配值的调控：与接受到高于社会平均分配值的分配（4/6 和 5/5）比，当接受到的分配值低于社会平均分配值时（1/9 和 2/8）被试产生了更大的 FRN。该实验发现**与预测 2 一致**。根据强化学习理论，FRN 表征消极的奖赏预测误差，实际反馈结果比预期越差时会诱发明显的 FRN（Holroyd & Coles, 2002）。我们的 FRN 结果表明大脑具有持续监测偏离社会规范的机制（Montague & Lohrenz, 2007），越是违反主观预期和社会规范的分配，越能诱发被试更大的 FRN。同时 FRN 结果也证明，本研究所采用的社会均值作为善意与否阈值的观点是正确的。

与预测 3a 一致，我们发现 P3 与分配结果的大小有关，低于社会均值的分配比高于社会均值的分配诱发被试更大的 P3。我们还发现真人分配比电脑分配诱发更大 P3 波幅，这与 P3 表征大脑对分配结果的主观价值评价过程（Gu et al., 2011; Yeung & Sanfey, 2004）有关。因为被试对真人分配具有更高的主观价值评价，消耗了更多的注意力资源，从而产生更大的 P3 波幅，**与预测 3b 一致**。更重要的是，本研究发现 P3 受到分配结果和意图交互效应的调控，**与预测 3c 一致**。该发现不但为我们的行为结果提供了直接的脑证据，还提示分配意图和分配结果可以相对独立地影响奖赏系统的神经活动（Bartholow et al., 2006）。

本文的核心贡献在于我们将社会均值视为上行间接互惠中判断 A 是否善意的参照点，精确刻画分配意图对上行间接互惠传递的重要作用，进而从行为和脑电两个层面实证检验了基于社会均值的上行间接互惠的传递机制。本文实验结果没有完全支持假说 1，但是我们通过将被试分为高低利他水平两组，分别考察他们的上行间接互惠行为特征发现，**高利他水平组的数据完全支持了假说 1**，而低利他水平组的被试没有表现出上述特征，这一发现表明个体利他水平异质性因素对是否传递高于社会均值的分配存在影响，后续研究可以在本文的研究基础上更全面地考察利他水平、关于社会分配均值的信念等异质性因素对上行间接互惠行为的影响。在 ERP 分析层面，后续研究在纳入更大样本量的典型高利他水平被试后，也应

重新考察这一因素对本文发现的 ERP 各指标的影响。

5 结论

本文首次创新性地考察了基于社会均值的上行间接互惠，采用行为结合 ERP 探讨了分配结果和分配意图对上行间接互惠行为的影响。结果发现，分配结果与意图均对上行间接互惠行为有显著的影响：与电脑的高于社会均值分配比，真人高于社会均值的分配让被试愿意分配给第三方更多；与电脑的低于社会均值分配比，真人低于社会均值的分配让被试分配给第三方更少。在脑电层面，早期加工阶段 N1 成分受到分配意图的调控，真人分配诱发更大的 N1；FRN 成分受到分配结果的调控，低于社会均值的分配比高于社会均值的分配诱发更大的 FRN；P3 成分受到分配结果和意图的交互调控。本研究从行为和脑电层面证明了分配结果和意图在上行间接互惠决策过程发挥着重要作用，为揭示消极和积极的间接互惠在什么条件下能够被抑制和激发提供了新的视角。

参考文献

Bartholow, B. D., Bushman, B. J., & Sestir, M. A. (2006). Chronic violent video game exposure and desensitization to violence: behavioral and event-related brain potential data. *Journal of Experimental Social Psychology, 42*(4), 532–539.

Blount, S. (1995). When social outcomes aren’t fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes, 63*(2), 131–144.

Boudreau, C., McCubbins, M. D., & Coulson, S. (2009). Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. *Social Cognitive and Affective Neuroscience, 4*(1), 23–34.

Chang, Y. -P., Lin, Y. -C., & Chen, L.H. (2012). Pay it forward: Gratitude in social networks. *Journal of Happiness Studies, 13*(5), 761−781.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 117*(3), 817−869.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology, 58*(6), 1015–1026.

Comesaña, M., Soares, A. P., Perea, M., Piñeiro, A. P., Fraga, I., & Pinheiro, A. (2013). ERP correlates of masked affective priming with emoticons. *Computers in Human Behavior, 29*(3), 588−595.

de Bruijn, E. R. A., Schubotz, R. I., & Ullsperger, M. (2007). An event-related potential study on the observation of erroneous everyday actions. *Cognitive, Affective, & Behavioral Neuroscience, 7*(4), 278−285.

Dong, G., Hu, Y., & Zhou, H. (2010). Event-related potential measures of the intending process: Time course and related ERP components. *Behavioral and Brain Functions, 6*(1), 15.

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology, 14*(5), 456–467.

Engel, C. (2011). Dictator game: A meta study. *Experimental Economics, 14*, 583–610.

Engelmann, D., & Fischbacher, U. (2009) . Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior, 67*(2), 399–407.

Falk, A., Fehr, E., & Fishbacher, U. (2008). Testing theories of fairness-Intentions matter. *Games and Economic Behavior, 62*(1), 287–303.

Falk, A., & Fishbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior, 54*(2), 293–315.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Fehr, E., Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*(3), 817–68.

Gehring, W. J. , & Willoughby, A. R. . (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295*(5563), 2279–2282.

Gong, Y., Yao, L,, Chen, X., Xia, Q., Jiang, J., & Du, X. (2022). Group membership modulates fairness consideration among deaf college students—An event-related potential study. *Frontiers in Psychology, 13*, 794892.

Gray, K., Ward, A. F., & Norton, M. I. (2014). Paying it forward: Generalized reciprocity and the limits of generosity. *Journal of Experimental Psychology: General, 143*(1), 247–254.

Gu, R., Lei, Z., Broster, L., Wu, T., Jiang, Y., & Luo, Y. J. (2011). Beyond valence and magnitude: A flexible evaluative coding system in the brain. *Neuropsychologia, 49*(14), 3891–3897.

Hayek, F. A. (1980). *Individualism and economic order (Reprinted.)*. Chicago London: the University of Chicago press.

Herrmann, C. S., Knight, R. T. (2001). Mechanisms of human attention: Event-related potentials and oscillations. *Neuroscience and Biobehavioral Reviews, 25*(6), 465–476.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679–709.

Horita, Y., Takezawa, M., Kinjo, T., Nakawake, Y., & Masuda, N. (2016). Transient nature of cooperation by pay-it-forward reciprocity. *Scientific Reports, 6*, 19471.

Hoy, C. W., Steiner, S. C., & Knight, R. T. (2021). Single-trial modeling separates multiple overlapping prediction errors during reward processing in human EEG. *Communications biology, 4*, 910.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J., & Weber, B. (2018). Spreading inequality: Neural computations underlying paying-it-forward reciprocity. *Social Cognitive and Affective Neuroscience, 13*(6), 578–589.

Huang, X., Li, J., & Ni, Y. (2023). Social norm modulates the enhancement effect of behavioral visibility on altruistic preference. *Acta Psychologica Sinica, 55*(3), 481–495.

[黄馨茹, 李健, 倪荫梅. (2023). 行为可见增加利他偏好及其社会规范机制. *心理学报, 55*(3), 481–495.]

Johnson, Jr., R., & Donchin, E. (1980). P300 and stimulus categorization: Two plus one is not so different from one plus one. *Psychophysiology, 17*(2), 167–178.

Kimbrough, E. O., & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association, 14*(3), 608–638.

Li, J., Pan, J., and Zhu, C. (2020). Inter-brain synchronization is weakened by the power to reject offers in bilateral bargaining games. *Social Cognitive and Affective Neuroscience, 17*(7), 625–633.

Lin, H. Y., Gao, H. W., You, J., Liang, J. F., Ma, J. P., Yang, N., et al. (2014). Larger N2 and smaller early contingent negative variation during the processing of uncertainty about future emotional events. *International Journal of Psychophysiology, 94*(3), 292–297.

Liu, M., Zhou, J., Liu, Y., & Liu, S. (2022). The impact of social comparison and (un)fairness on upstream indirect reciprocity: Evidence from ERP. *Neuropsychologia, 177*, 108398.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique.* 2nd Edn. Cambridge, MA: MIT press.

Ma, Q., & Hu, Y. (2015). Beauty Matters: Social Preferences in a Three-Person Ultimatum Game. *PLoS ONE, 10*(5): e0125806.

Ma, Q., Hu, Y., Jiang, S., & Meng, L. (2015). The undermining effect of facial attractiveness on brain responses to fairness in the Ultimatum Game: an ERP study. *Frontiers in Neuroscience, 9*, 77.

Mayer, S. V., Rauss, K., Pourtois, G., Jusyte, A., & Schönenberg, M. (2019). Behavioral and electrophysiological responses to fairness norm violations in antisocial offenders. *European Archives of Psychiatry and Clinical Neuroscience, 269*, 731–740.

McBride, M., & Ridinger, G. (2021). Beliefs also make social-norm preferences social. *Journal of Economic Behavior & Organization, 191*(3), 765–784.

McCullough, M. E., Kilpatrick, S. D., Emmons, R. A., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin, 127*(2), 249.

Messick, D. M., & Schell, T. (1992). Evidence for an equality heuristic in social decision making. *Acta Psychologica, 80*(1-3), 311–323.

Miraghaie, A. M., Pouretemad, H., Villa, A. E. P., Mazaheri, M. A., Khosrowabadi, R., & Lintas, A. (2022). Electrophysiological Markers of Fairness and Selfishness Revealed by a Combination of Dictator and Ultimatum Games. *Frontiers in Systems Neuroscience, 16*, 765720.

Montague, P. R., & Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. *Neuron, 56*(1), 14–18.

Moore, M., Katsumi, Y., Dolcos, S., & Dolcos, F. (2021). Electrophysiological Correlates of Social Decision-making:An EEG Investigation of a Modified Ultimatum Game. *Journal of Cognitive Neuroscience, 34*(1), 54–78.

Niemand, T, Mai R, & Kraus S. (2019). The zero-price effect in freemium business models: The moderating effects of free mentality and price-quality inference. *Psychology & Marketing, 36*(8), 773–790.

Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology, 194*(4), 561–574.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*, 1291–1298.

Pereda, M., Brañas-Garza, P., Rodríguez-Lara, I., & Sánchez, A. (2017). The emergence of altruism as a social norm. *Scientific Reports, 7*(1), 9684.

Qu, C., Wang, Y., & Huang, Y. (2013). Social exclusion modulates fairness consideration in the ultimatum game: An ERP study. *Frontiers in Human Neuroscience, 7*, 505.

Rutte, C., & Taborsky, M. (2007). Generalized Reciprocity in Rats. *PLoS Biology, 5*(7), e196.

Sigmund, K. (2010). *The Calculus of Selfishness*. Princeton University Press, Princeton.

Stanca, L., Bruni, L., & Corazzini, L. (2009). Testing theories of reciprocity: Do motivations matter? *Journal of Economic Behavior & Organization, 71*(2), 233–245.

Sun, Y. X., Zhang, J. H., & Li, J. P. (2022). Research progress on indirect reciprocity. *Economic Perspectives,* (1), 146–160.

[孙熠譞, 张建华, 李菁萍. (2022). 间接互惠理论研究进展. *经济学动态,* (1), 146–160.]

Wei, H., & Zhou, R. (2020). High working memory load impairs selective attention: EEG signatures. *Psychophysiology, 57*(11), e13643.

Wu, Y., Leliveld, M. C., & Zhou, X. (2011). Social distance modulates recipient’s fairness consideration in the dictator game: An ERP study. *Biological Psychology, 88*(2-3), 253–262.

Xie, H., Hu, X., Mo, L., & Zhang, D. (2021). Forgetting positive social feedback is difficult: ERP evidence in a directed forgetting paradigm. *Psychophysiology, 58*(5), e13790.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of Neuroscience, 24*(28), 6258–6264.

Zhong, X, Wang R, Huang S, Chen J, Chen, H, & Qu, C. (2019). The neural correlate of mid-value offers in ultimatum game. *PLoS ONE, 14*(8), e0220622.

Zhang, D. D., Gu, R., Wu, T., Broster, L. S., Luo, Y., Yang, J., & Luo, Y. (2013). An electrophysiological index of changes in risk decision-making strategies. *Neuropsychologia, 51*(8), 1397–1407.

Zhang, Y., Yu, H., Yin, Y., & Zhou, X. (2016). Intention Modulates the Effect of Punishment Threat in Norm Enforcement via the Lateral Orbitofrontal Cortex. *The Journal of Neuroscience, 36*(35), 9217–9226.

Intention and Upstream indirect reciprocity: Insights from behavioral and ERP evidence

WANG Ting[1]; ZHAO Liangfo[2]; YANG Jinpeng[2]; ZHANG Dandan[2,3]; LEI Zhen[2,4]

(1 Business School, Sichuan University, Chengdu 610065, China;

2 China Center for Behavioral Economics and Finance, Southwestern University of Finance and Economics, Chengdu 611130, China;

3 Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu 610066, China;

4 School of Economics , Southwestern University of Finance and Economics, Chengdu 611130, China)

Abstract

Upstream indirect reciprocity, a widespread phenomenon observed both in real-world settings and controlled experimental environments, extends beyond conventional reciprocity systems and plays a crucial role in fostering large-scale human cooperation and maintaining social order. Although this social phenomenon has garnered significant scholarly attention, existing research remains insufficient in uncovering its underlying mechanisms.

Previous studies typically use a two-stage dictator game to investigate upstream indirect reciprocity. According to meta-analyses of experimental literature on dictator games, dictators typically allocated 28% of the total to recipients, with allocations exceeding 50% being extremely rare (Engel, 2011). Surprisingly, most existing research considers equal distribution (where A allocates 50% to B) as the threshold for determining whether A has good intentions (e.g., Gray et al., 2014; Hu et al., 2018), despite lacking sufficient justification for this criterion. Different from prior work, we propose that individuals assess others' intentions relative to the social mean as a reference point. Based on this premise, we hypothesize that when individuals receive an allocation above the social mean, they are more likely to pass on a value above the mean to third parties, whereas allocations below the mean will result in values passed below the mean. If individuals indeed pass on a higher or lower value based on whether they received above- or below-mean allocations, this result might also be explained by an alternative hypothesis: the income effect, where people give more when they have more resources. Therefore, this study investigates whether intention-based indirect reciprocity persists event after controlling for the income effect.

We recruited 42 undergraduate participants for the experiment, which consists of two parts: a standard dictator game followed by an indirect reciprocity experiment. The second experiment employed a 2 (distribution below the social mean vs. above the social mean) × 2 (human allocation vs. computer allocation) within-subject design. The main experiment featured both a human allocation task and a computer allocation task, with task order counterbalanced among participants. Each task included 156 trials, for a total of 312 trials.

The results show that both distribution outcomes and perceived intentions significantly influence upstream indirect reciprocity. Specifically, participants allocated more to third parties after receiving above-mean distributions from a human compared to a computer, while

below-mean distributions from a human led to lower allocations than those from a computer. EEG data revealed that N1 components were modulated by perceived intention, with human allocations eliciting greater N1 responses. The feedback-related negativity (FRN) component was influenced by distribution outcomes, with below-mean distributions evoking larger FRN responses than above-mean distributions. Finally, the P3 component was regulated by the interaction between distribution outcome and intention.

Key words: upstream indirect reciprocity, social norms, intention, event-related potential